A new and intuitive test for zero–modification

Paul Wilson [1] and Jochen Einbeck [2]

[1] School of Mathematics and Computer Science, University of Wolverhampton, United Kingdom

[2] Department of Mathematical Sciences, Durham University, United Kingdom

---

Address for correspondence: Paul Wilson, School of Mathematics and Computer Science, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1LY, United Kingdom .

E-mail: pauljwilson@wlv.ac.uk.

Phone: (+44) 1902 321 444.

Fax: (+44) 1902 321 478.

Abstract: While there do exist several statistical tests for detecting zero–modification in count data regression models, these rely on asymptotical results and do not transparently distinguish between zero–inflation and zero–deflation. In this manuscript, a novel non–asymptotic test is introduced which makes direct use of the fact that the distribution of the number of zeros under the null hypothesis of no zero–modification can be described by a Poisson–binomial distribution. The computation of critical values from this distribution requires estimation of the mean parameter under the null hypothesis, for which a hybrid estimator involving a zero–truncated mean estimator is proposed. Power and nominal level attainment rates of the new test are studied, which turn out to be very competitive to those of the likelihood ratio test. Illustrative

data examples are provided.

Key words: Zero–inflation, count data, Poisson–binomial distribution, Zero–truncated Poisson distribution, mid–p–values

1 Introduction

There are many reasons why one would observe (or suspect) that the number of zeros in a given count data set is unusually large or unusually small. These reasons can be roughly classified into two major categories: (i) Bias arising from the data collection procedure (ii) Structural zeros due to an underlying physical reason. To give an example for (i), we cite Dietz and Böhning (2000) who modelled zero–deflated DMFT index data from a dental epidemiological study previously published by Mendonca (1995). Specifically, the DMFT index quantifies the dental status of an individual through a count of "Decayed, Missing and Filled Teeth", and it was noted that an "incorrect sampling procedure" had led to the non–inclusion of some children whose score was zero.

An example for (ii) is illustrated through the two data sets displayed in Table 1, which report results from laboratory (*in vitro*) experiments where frequencies of chromosome aberrations were counted after exposing blood samples to 200 kV X–rays (Heimers et al., 2006). To be more precise, blood (from healthy volunteers) was mixed and then divided into five parts, with each part getting exposed to one of the doses $1\text{Gy}, \ldots, 5\text{Gy}$. The radiation exposure may lead to double–strand breaks, which, when incorrectly repaired by the DNA–damage response mechanism, can produce dicentric chromo-

somes (that is, chromosomes with two centromeres) or centric rings, which can be counted under a microscope. While Table 1 (left) is representing data collected under a 'whole–body–exposure' scenario, Table 1 (right) represents a partial exposure scenario in which 25% exposed blood was mixed with 75% unexposed blood. It is clear that the three quarters of blood which have not been exposed to radiation will contribute very little chromosome aberrations (there does exist a background prevalence of such aberrations, for instance caused by naturally occurring ionizing radiation, but this rate is very low). Hence, one naturally would assume many 'structural' zeros in this data set, as is indeed observed.

Such considerations lead to the question of what it actually means to speak of 'too few' or 'too many' zeros. Usually this notion is related to a specific statistical model. For instance, in the field of radiation biodosimetry, the model of choice has been traditionally the Poisson model, based on solid physical arguments and empirical evidence. If the number of zeros is too large or too small relative to what would be expected under the assumed model, be it due to bias or for structural reasons, the Poisson model will fit poorly. A possible solution to the problem is to resort to a more complex model. In the case of partial body radiation exposure, a zero–inflated model appears to be a natural choice, though a plethora of alternative models including the negative binomial distribution and Hermite models have been suggested for this kind of data (Oliveira et al., 2016).

But, taking the decision aside on which alternative model to choose, it remains the immediate question of whether or not there is evidence for deflation or inflation of zeros relative to the baseline model. While this seems quite likely in Table 1, where the right hand table features much more zeros (and far fewer ones) than the left

Table 1: Number of chromosome aberrations in blood samples exposed to sparsely ionizing radiation. Left: whole body exposure scenario; right: partial body exposure scenario. These data sets have been labelled (A3) and (C1) in Oliveira et al. (2016).

(A3)	frequency							
dose	0	1	2	3	4	5	6	7
1	1715	268	15	2	0	0	0	0
2	638	298	56	8	0	0	0	0
3	247	225	85	37	6	0	0	0
4	99	129	92	52	21	5	2	0
5	48	88	97	99	36	25	5	2

(C1)	frequency								
dose	0	1	2	3	4	5	6	7	8
1	2713	78	8	0	1	0	0	0	0
2	1302	71	22	5	0	0	0	0	0
3	1116	46	28	7	2	1	0	0	0
4	929	18	14	22	13	2	0	1	1
5	726	17	18	12	9	13	1	4	0

hand table, this question would be much harder to assess if we hadn't seen the left hand side table. Hence, there is a need for quantitative methods which, relative to a given model, help to decide whether zero–inflation or deflation exists. The terms zero–inflation and zero–deflation have sometimes been combined towards zero–modification, meaning that there are either too few or too many zeros in the data, relative to the specified count data model. We follow this convention henceforth.

Of course, such methods do exist, in principle, already in the statistician's toolbox, with the most prominent representative being the likelihood ratio test, which we will outline in detail in Section 2.2. Also score (Rao) and Wald tests are available for this purpose. While these tests are all viable, they rely upon asymptotic results and hence implicitly on large samples, and they do, in their standard form, not transparently distinguish between zero–inflation and zero–deflation (at least not without proper adjustment, which will be unknown to many applied users). It should also be noted that, whilst Vuong's test for non-nested models has recently become popular as a test for zero–inflation, Wilson (2015) shows such use to be methodologically erroneous.

We propose here a new and intuitive test of zero–modification that avoids such issues elegantly and which possesses similar attainment rates and power to previous tests. The proposed test relates more directly to the character of zero–modification than the other tests: the test will employ the number of zeros in the data as the test statistic, and tests whether this number is consistent with the non zero–modified model. We will demonstrate that this statistic, under the null hypothesis of no zero–modification, follows a Poisson–binomial distribution, based on which critical values can be obtained.

The rest of the paper develops as follows. Section 2 will introduce the test problem, and review the likelihood ratio test in this context. Section 3 will introduce our new test of zero–modification. Section 4 will discuss the important question of how to robustly estimate the Poisson mean parameter (which is needed for the computation of our test statistic) in the absence of the knowledge of whether or not the Poisson assumption is correct. Section 5 provides real data examples with and without covariates, including a detailed study of the chromosome aberration data. Section 6 provides concluding remarks.

2 Testing for zero-modification

2.1 Hypotheses

To fix terms, denote $y = \{y_1, y_2, \ldots, y_n\}$ an independent sample drawn from count random variables $Y_1, \ldots, Y_n$. We denote further the mean of y by $\bar{y}$, and the number of zeros in y by n_0, which can be considered as a realization of the random variable

$N_0 = \sum_{i=1}^{n} \mathbf{1}_{\{Y_i=0\}}$.

The question of interest is whether the distribution of the Y_i is zero–modified with respect to a given count distribution $F(y_i|\mu_i, \phi)$ with densities $p(y_i|\mu_i, \phi)$, where the mean parameter μ_i may depend on a set of covariates $\boldsymbol{x}_i \in \mathbb{R}^d$, $i = 1, \ldots, n$ in some pre–specified form, and ϕ captures further model parameters such as shape or scale parameters of F. (In principle ϕ could be modelled by further covariates though for ease of presentation we assume that this is not the case.) That is, we assume that $g(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}$ with some monotonic and known link function g, and model parameters $\boldsymbol{\beta} \in \mathbb{R}^d$ which will have to be estimated. While the new test procedure is applicable to test for zero–modification w.r.t *any* baseline count distribution, in this work the most important application will be the Poisson distribution, in which case μ_i corresponds just to the Poisson parameter, and ϕ is empty.

Expressing the general framework above in other words, we wish to establish whether the distributional assumption $Y_i|\boldsymbol{x}_i \sim F(y_i|\mu_i, \phi)$ is consistent with the number of zeros observed. It is clear that both the count distribution F and the predictor specification for μ_i impact on the model fit. We consider our test as a tool to asses the adequacy of F given the specification of μ_i, but not as a tool to simultaneously assess F and μ_i. Hence, we use F in what follows as short hand notation for the entire model specification, that is we identify notationally $F \equiv F(y_i|\mu_i, \phi)$.

We formulate the null hypotheses and three possible alternatives as follows:

$$\begin{aligned} H_0 &: \text{The distribution of } Y_i|\boldsymbol{x}_i \text{ follows the specified count data model } F. \\ H_1^{(a)} &: \text{The distribution of } Y_i|\boldsymbol{x}_i \text{ is zero–modified w.r.t count data model } F. \\ H_1^{(b)} &: \text{The distribution of } Y_i|\boldsymbol{x}_i \text{ is zero–inflated w.r.t count data model } F. \\ H_1^{(c)} &: \text{The distribution of } Y_i|\boldsymbol{x}_i \text{ is zero–deflated w.r.t count data model } F. \end{aligned} \tag{2.1}$$

Notably, our approach will not require fitting the model under the alternative, which is a property shared with the score test but not with the Wald test and the likelihood ratio test. The latter procedure, which can be considered as the most prominent among the three asymptotic sister tests, is briefly reviewed below.

2.2 Likelihood ratio tests

Likelihood ratio tests are usually employed to determine whether a larger model fits significantly better than a competing smaller (or 'restricted') model that is nested within it (though some variants for non–nested models have also been proposed, see for example Cox (1962) and Vuong (1989)). In the case of testing for zero–inflation, the larger model takes the shape

$$p(y_i|\mu_i, \phi, \omega) = (1 - \omega)p(y_i|\mu_i, \phi) + \omega p(y_i, 0) \tag{2.2}$$

where $\omega \geq 0$ is the zero–inflation parameter, $p(y_i|\mu_i, \phi)$ is some base density (corresponding to the restricted model) such as Poisson or Negative binomial, and $p(y_i, 0)$ is a point mass at 0, that is $p(y_i, 0) = \mathbf{1}_{\{y_i=0\}}$. The test problem of zero–inflation can then be stated as

$$\tilde{H}_0 : \quad \omega = 0 \quad \text{versus} \quad \tilde{H}_1 : \quad \omega > 0.$$

For nested models where the smaller model does not sit on the boundary of the parameter space of the larger model, it is well known that the distribution of the LR test statistic (under the restricted model, corresponding to the null hypothesis) follows a χ^2 distribution, with the degrees of freedom being equal to the number of parameters by which the two models differ. However, when testing for zero–inflation then we are precisely in a scenario where the restricted model, for $\omega = 0$, does sit on that boundary. Molenberghs and Verbeke (2007) showed that the resulting LR test statistic

$$\varrho = -2\left[\ell\left(\hat{\boldsymbol{\beta}}^{(r)}, \hat{\phi}^{(r)}, 0\right) - \ell\left(\hat{\boldsymbol{\beta}}, \hat{\phi}, \hat{\omega}\right)\right], \tag{2.3}$$

where $\ell(\boldsymbol{\beta}, \phi, \omega) = \sum_{i=1}^{n} \log p(y_i|\mu_i, \phi, \omega)$, with $\mu_i = g^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta})$, and the superscript (r) indicating that all model parameters have been estimated under the restriction $\omega = 0$, follows an equal mixture of a χ^2_0 (i.e. a point mass at zero) and a χ^2_1 distribution. Table 2 compares the theoretical 95%, 98% and 99% quantiles of such a distribution with the estimates of those quantiles of the distribution of the log–likelihood ratios (based upon 10000 resamples) when zero–*inflated* Poisson and Poisson models are fitted to samples of sizes 1000, 40 and 20 drawn from Poisson data with parameters $\mu = 2$, $\mu = 0.8$ and $\mu = 0.5$ respectively. As is apparent, even for relatively large sample sizes and Poisson means, the approximation is somewhat poor.

While model (2.2) was originally only thought as a 'zero–inflated' model, it actually allows for zero–deflation. In the particular case of zero–modified Poisson (ZMP), where $p(y_i|\mu_i) = e^{-\mu_i}\mu_i^{y_i}/y_i!$, one can show that the density is still well–defined for all $\omega \geq -\frac{1}{e^{\mu_i}-1}$. The test problem of zero–modification can then be stated as

$$\tilde{H}_0: \quad \omega = 0 \quad \text{versus} \quad \tilde{H}_1: \quad \omega \neq 0.$$

and in this case the asymptotic distribution reverts to χ^2_1. However, note that espe-

Table 2: Observed quantiles of the distribution of the log-likelihood ratios ZIP versus Pois(μ) and the theoretical quantiles under a $0.5\chi^2_0 + 0.5\chi^2_1$ distribution

		quantiles		
n	μ	95%	98%	99%
1000	2.0	2.592	4.099	5.461
40	0.8	2.340	3.758	5.082
20	0.5	2.235	4.415	4.785
theoretical		2.706	4.218	5.412

cially (but not only) in the presence of covariates, often a monotonic link function r is applied, with common choices being the complementary log–log (cloglog) link $r(\omega) = \log(-\log(1-\omega))$ or the logit link, $r(\omega) = \log(\omega/(1-\omega))$. Notably, the use of a logit or cloglog link excludes the detection of zero–deflation since they imply the restriction $\omega > 0$. Hence, if zero–deflation is to be detected then the identity link $r(\omega) = \omega$ is the best choice. It is finally noted, that, even though the likelihood ratio test can be used to test for any of zero–inflation, zero–deflation, or zero–modification in principle, the LR test statistic (2.3) as such is uninformative for the direction of the modification.

3 The proposed test

3.1 Distribution of test statistic

Assume H_0 is true and hence $F = F(y_i|\mu_i, \phi)$ is the correct model, with density $p(y_i|\mu_i, \phi)$. Let $p_i = p(0|\mu_i, \phi)$ (that is, in the Poisson case, $p_i = e^{-\mu_i}$), and T_i a

random variable which takes the value 1 if $Y_i = 0$ and 0 otherwise. Clearly, T_i is a Bernoulli random variable with parameter p_i, and so the random variable N_0, which serves as test statistic, can be formulated as the sum over independent Bernoulli experiments $T_1, \ldots, T_n$.

Based on this simple observation, consider the special case that there are no covariates, that is $\mu_1 = \mu_2 = \cdots = \mu_n = \mu$. In this case, the p_i's are equal also, and so the distribution of N_0 is the binomial distribution $\text{Bin}(n, p)$, and thus has mean np and variance $np(1-p)$. Based on this distribution, one can immediately compute quantiles corresponding to a given significance level, and use these as critical values for the test; see Section 3.2.

The situation is more interesting when μ_i *does* depend on covariates $\boldsymbol{x}_i$, that is $\mu_i = g^{-1}(\boldsymbol{x}_i^T \boldsymbol{\beta})$, and hence the p_i's are not all equal. The distribution of a sum of Bernoulli distributions with different success probabilities is known as a *Poisson–binomial* distribution (Chen and Liu, 1997), with probability mass function

$$P(N_0 = k) = \left\{ \prod_{i=1}^{n} (1 - p_i) \right\} \sum_{i_1 < \cdots < i_k} w_{i_1} \cdots w_{i_k}$$

where $w_i = \frac{p_i}{1-p_i}$, $i = 1, 2, \ldots, n$, and the summation is over all possible combinations of distinct $i_1, i_2, \ldots, i_k$ from $\{1, 2, \ldots, n\}$.

Note that this is not a *compound* Poisson–binomial distribution. Daskalakis et al. (2012) remark that "It is believed that Poisson (1837) was the first to consider this extension of the binomial distribution, and the distribution is sometimes referred to as 'Poisson's binomial distribution' ".

The R package `poibin` (Hong, 2013b) implements both exact and approximate methods for computing the cumulative distribution function of the Poisson–binomial dis-

tribution based upon algorithms presented by Hong (2013). It also provides the probability mass function, quantile function and random number generation for the Poisson–binomial distribution. Four options for the model fitting algorithms are available in `poibin`, throughout this paper we use the default *DFT-CF* algorithm.

3.2 Test procedure

To carry out the actual test, specify a significance level α, and decide for one of the test scenarios (a), (b) or (c) as given in (2.1). Denote by n_γ an appropriate γ–quantile of the Poisson-binomial distribution of N_0 (to be discussed below). The test consists of carrying out the following procedure:

(i) Fit the relevant count data regression model to the data, yielding means

$$\hat{\mu}_i = g^{-1}(\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}), \tag{3.1}$$

and, if relevant, further distributional parameters $\hat{\phi}$;

(ii) for each y_i estimate $\hat{p}_i = p(0|\hat{\mu}_i, \hat{\phi})$;

(iii) use a Poisson–binomial distribution with parameters $(\hat{p}_1, \ldots, \hat{p}_n)$ to determine the distribution of N_0. [This reduces to the binomial distribution $\text{Bin}(n, \hat{p})$ in the absence of covariates, where $\hat{p} \equiv \hat{p}_1 = \cdots = \hat{p}_n$.]

(iv) Depending on the chosen alternative, do one of the following

(a) Reject H_0 in favour of $H_1^{(a)}$ if $n_0 < n_{\alpha/2}$ or $n_0 > n_{1-\alpha/2}$.

(b) Reject H_0 in favour of $H_1^{(b)}$ if $n_0 > n_{1-\alpha}$.

(c) Reject H_0 in favour of $H_1^{(c)}$ if $n_0 < n_\alpha$.

Otherwise, one fails to reject H_0.

For the use in our test, appropriate quantiles, or, equivalently, p–values, need to be extracted from the relevant Poisson-binomial distribution. For instance, for test problem (b), the customarily defined quantile and p-value are given by $n_{1-\alpha} = \min_t\{P(N_0 \leq t) \geq 1-\alpha\}$ and $p^*(t) = P[N_0 \geq t]$, respectively. However, it has been argued in the literature that these quantities behave unfavourably for discrete distributions, both from a theoretical and practical viewpoint (Ma et al., 2011; Franck, 1986). The latter reference strongly advocates the use of the *mid-p-value*, drawing on previous research by Lancaster (1961), Dempster and Schatzoff (1965) and Stone (1969). Specifically, for a given value t of the test statistic N_0, the mid–p–value is given by

$$p^*_{0.5}(t) = P[N_0 > t] + 0.5P[N_0 = t] = 0.5\left(P\left[N_0 \geq t\right] + P[N_0 \geq t+1]\right), \qquad (3.2)$$

and, under the null hypothesis, enjoys the property that $E(p^*_{0.5}(N_0)) = 0.5$ unlike for the customarily defined p–value for which this expectation may range between $1/2$ and 1 for discrete distributions (Franck, 1986).

Following similar lines of reasoning, one can motivate and define the *mid–quantile* (Ma et al., 2011); the mathematical definition of which is a bit lengthy and is therefore omitted here. For a precise formulation, in the context of the test under study, see Wilson and Einbeck (2017). For all application studies to be carried out in Section 5, we will employ mid–p–values and mid–quantiles. We refer to the interval $[n_{\alpha/2}, n_{1-\alpha/2}]$ as a $1-\alpha$ mid–quantile interval (MQI) for N_0.

4 Estimating the Poisson parameter

A key component of our test which has not been discussed in detail yet is how to estimate the mean function (3.1) in step (i) of the test introduced in Subsection 3.2. The reader may be surprised that there is an issue at this stage — the problem is that we need to estimate a Poisson mean parameter in the absence of the knowledge of whether this Poisson assumption is correct, that is whether there is zero–modification or not. For a given sample $y = \{y_1, y_2, \ldots, y_n\}$ without covariates, the 'obvious' choice under the Poisson assumption would be the 'whole sample mean' $\hat{\mu}_{\mathrm{W}} = \bar{y}$, which corresponds to the maximum likelihood estimator, and is unbiased for μ. However, as we will demonstrate in Subsection 4.2, this estimate may lead to a severe underestimation of μ if the data is in fact zero–inflated, or an overestimation if the data is in fact zero–deflated. We therefore consider in Section 4.1 an alternative mean estimator, based on the zero–truncated distribution, which resolves this problem at the expense of an increased variance. A hybrid version of the two estimators is introduced in Subsection 4.3 and its properties in terms of the test under consideration are analyzed in Subsection 4.4. For ease of presentation, all considerations in Subsections 4.1 to 4.4 are provided in the case without covariates. The required adaptations when including covariates are discussed in Subsection 4.5.

4.1 Estimation through zero–truncated distribution

As before we denote by n_0 the number of zero–valued observations in y. When the latter is Poisson, then the distribution of the $n - n_0$ non–zero observations will follow

a zero–truncated Poisson distribution $Z \sim ZTP(\mu)$ with probability mass function

$$p_{\mathrm{T}}(z|\mu) = \frac{\mu^z}{(e^\mu - 1)z!} \tag{4.1}$$

for $z = 1, 2, 3, \ldots$. It is well known that

$$\zeta \equiv E(Z) = \frac{\mu e^\mu}{e^\mu - 1} \equiv s(\mu) \tag{4.2}$$

and hence

$$\mu \equiv s^{-1}(\zeta). \tag{4.3}$$

Irwin (1959) gives an explicit expression for $s^{-1}(\zeta)$ involving a Lagrange series expansion, this is sometimes slow to converge and not conveniently implementable. Plackett (1953) shows that μ can be estimated without bias through the expression $\sum_{y_i \geq 2} y_i/(n - n_0)$. Ridout and Demétrio (1992) show that a very accurate estimate of μ may be obtained using

$$\hat{\mu}_{\mathrm{T}} = \frac{\hat{\zeta}\left[1 - \exp\left(-s_1(\hat{\zeta})\right)\right]^2 - \left[s_1(\hat{\zeta})\right]^2 \exp\left[-s_1(\hat{\zeta})\right]}{1 - \left[s_1(\hat{\zeta}) + 1\right] \exp\left[-s_1(\hat{\zeta})\right]} \tag{4.4}$$

where

$$s_1(\hat{\zeta}) = \hat{\zeta}\left[1 - \exp\left(\frac{1}{\hat{\zeta}} - \hat{\zeta}\right)\right] \tag{4.5}$$

and $\hat{\zeta}$ is the mean of the positive observed data. We use estimator (4.4) in what follows.

4.2 Bias and precision of estimators

For $Y_i \sim \mathrm{Pois}(\mu)$, it is important to recognise that whilst, unconditionally, $\hat{\mu}_W$ is an unbiased estimator of μ, this is not the case when conditioning on the number of observed zeros, n_0. With N_0 as defined in Subsection 2.1, and assuming w.l.o.g. that

the first observations $Y_1, \ldots, Y_{n-n_0}$ give the non–zero results, one has from (4.2)

$$\begin{aligned} E\left(\hat{\mu}_{\mathrm{W}} | N_0 = n_0\right) &= \frac{1}{n} \sum_{i=1}^{n-n_0} E(Y_i | Y_i > 0) \\ &= \frac{n - n_0}{n} \frac{\mu e^{\mu}}{e^{\mu} - 1} \\ &= \left(1 - \frac{n_0}{n}\right) \frac{e^{\mu}}{e^{\mu} - 1} \cdot \mu. \end{aligned} \tag{4.6}$$

If we substitute $ne^{-\mu}$ (i.e. $E(N_0)$) for n_0 in (4.6) the right-hand side reduces to μ, and hence if $n_0 > E(N_0)$ the Poisson parameter tends to be underestimated, and if $n_0 < E(N_0)$ it tends to be overestimated. It is worth noting that the derivation of (4.6) remains valid when allowing for zero–modification (that is when assuming the Poisson assumption to hold only for the non–zero part).

In contrast, the estimator $\hat{\mu}_{\mathrm{T}}$ of (4.3) does not incur bias when conditioning on n_0, since the number of zeros is not involved in its calculation. However, it is less precise than $\hat{\mu}_W$. This is illustrated in Figure 1 which shows the estimates of the Poisson means obtained when $n = 100$ observations are sampled from a Pois(1) distribution. The black circles indicate whole sample mean (Poisson) estimates $\hat{\mu}_{\mathrm{W}}$, and the grey crosses the estimates $\hat{\mu}_{\mathrm{T}}$ obtained from the positive observations. The horizontal axis gives the number of zeros, n_0, with the expected number of zeros under the Poisson model, $100e^{-1} \approx 37$, highlighted by a dotted line. It is clear that the whole sample mean estimator has smaller variance but is biased if the observed number of zeros is far from their expected number. On the other hand, the ZTP–derived mean estimator does not demonstrate a noticeable bias, at the expense of a large variance.

The unsuitability of using either $\hat{\mu}_{\mathrm{W}}$ or $\hat{\mu}_{\mathrm{T}}$ in our test problem is shown by Figure 2. The left hand diagram illustrates the attainment of the test for H_0 vs $H_1^{(a)}$ with a nominal significance level of 0.05 for sample of size $n = 500$ using $\hat{\mu}_{\mathrm{W}}$, and the right

Figure 1: Estimation from the zero–truncated and whole sample

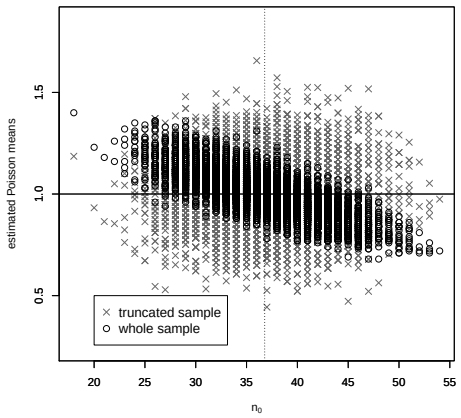
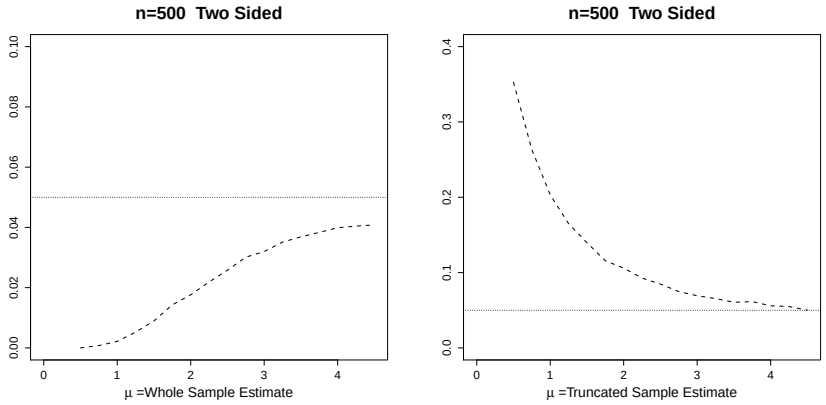


Figure 2: Attainment rate using $\hat{\mu}_{\mathrm{W}}$ and $\hat{\mu}_{\mathrm{T}}$



hand diagram using $\hat{\mu}_{\mathrm{T}}$. Clearly, even for such a sample size, neither estimator is suitable.

4.3 A hybrid estimator

We propose here a hybrid estimator for the Poisson parameter, μ, that balances the precision of $\hat{\mu}_{\mathrm{W}}$ with the accuracy of $\hat{\mu}_{\mathrm{T}}$:

$$\hat{\mu}_{\mathrm{H}} = h\hat{\mu}_{\mathrm{W}} + (1-h)\hat{\mu}_{\mathrm{T}} \qquad 0 \leq h \leq 1. \tag{4.7}$$

Iterative schemes which alternately optimize h (in terms of MSE) and update $\hat{\mu}_{\mathrm{H}}$ were considered, but found rather unsuitable since the additional variance created in this process defeats the purpose of the hybrid estimator. Instead, we give the following, simpler, recommendations based on simulation studies which are presented in summarized form in Figure 3. It is apparent that for larger mean parameter values the value of h is less critical than for smaller values, and that

(i) $h = 2/3$

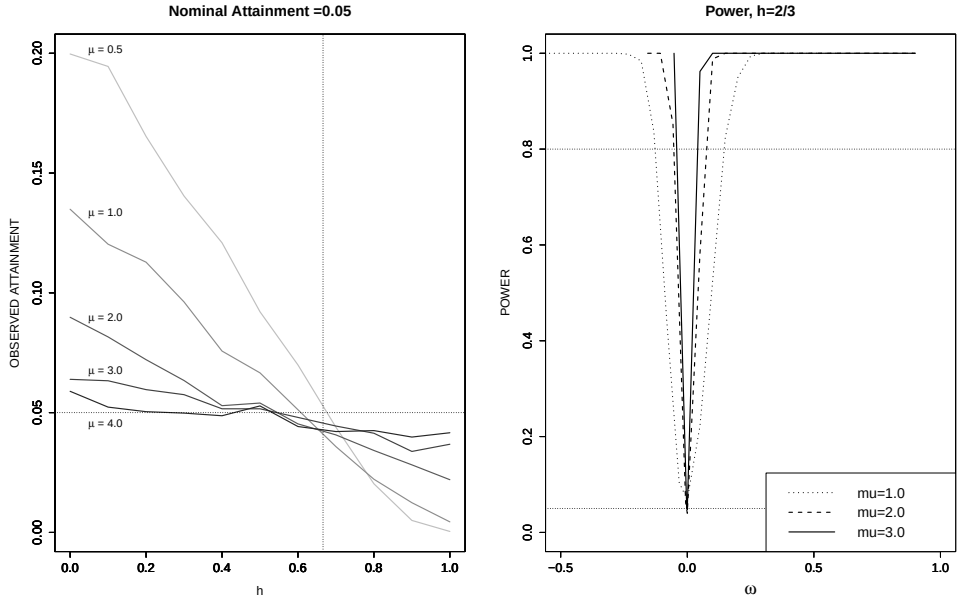
returns a parameter estimate that results in good power and attainment of the nominal level of significance for all values of the Poisson parameter. Based on comprehensive simulations which we have carried out but do not present in detail, we also suggest an 'adaptive' selection method for h, that results in slightly improved power and attainment, namely

(ii)

$$h = f(\hat{\mu}_{\mathrm{W}}) = \begin{cases} 0.7\left(0.85^{\hat{\mu}_{\mathrm{W}}}\right) & \hat{\mu}_{\mathrm{W}} < \frac{\ln(5/7)}{\ln(17/20)} \\ \frac{1}{2} & \text{otherwise,} \end{cases} \tag{4.8}$$

the rationale for which being that smaller mean parameters will lead to many zeros and thus few positive observations, hence the weight of the truncated estimator should decrease in this case. (The constant $\frac{\ln(5/7)}{\ln(17/20)} \approx 2.07$ is chosen so that f is continuous). Detailed study of the performance of schemes (i) and (ii), for one–sided and two–sided tests, is provided as follows.

Figure 3: Left: observed attainment under various values of h; right: observed power for $h = 2/3$.



4.4 Attainment and power of the proposed test

In a simulation study, the nominal level attainment of the proposed tests was studied under a nominal 0.05 level of significance and sample sizes of 500, 100 and 30. Figure 4 shows the attainment rates as a function of the true Poisson parameter μ for both schemes, 'fixed' and 'adaptive', with the corresponding rates for the likelihood ratio test shown for comparative purposes. Results for both a two tailed test of zero–modification (alternative hypothesis $H_1^{(a)}$) and a one tailed test of zero–inflation ($H_1^{(b)}$) are presented. It is apparent that, for both test scenarios, both the fixed and adaptive mixing parameters have excellent attainment rates, the latter especially so.

Figures 5 and 6 show the power of the proposed zero–modification and zero–inflation tests, respectively, for sample sizes of 500, 100 and 30 and Poisson parameters of 0.5, 1 and 2. The power of the likelihood ratio test is shown for comparative purposes. It

is observed that, under both test scenarios, the adaptive and fixed mixing parameters lead to tests with nearly identical powers which are either extremely similar to that of the likelihood ratio test, or greater. The relatively weaker power of the LR test becomes more pronounced for small Poisson parameters and sample sizes, noting however that for very small sample sizes the comparison becomes difficult since then all attainment curves behave rather erratically (Figure 4 bottom).

Concerning the execution of the simulation, in the right-hand side diagrams of Figure 4 and the diagrams of Figure 6, which pertain to the one–sided version of the test, the competing models of the likelihood ratio test are a Poisson model, where μ is modelled by a log link, and a zero–inflated Poisson model, where μ is modelled by a log link and ω by a logit link. For the estimates $\hat{p}_i = p\left(0|\hat{\mu}_i\right) = e^{-\hat{\mu}_i}$ required for the proposed test, the estimate of μ also uses a log link. In the left-hand side diagrams of Figure 4 and the diagrams of Figure 5, which pertain to the two–sided version of the test, the competing models of the likelihood ratio test are a Poisson model, where μ is modelled by an identity link, and a zero–modified Poisson model, where both μ and ω are modelled by an identity link, and the estimates of μ required for the proposed test are also derived using identity links. Depending upon the value of the Poisson parameter and the sample size either 5000 or 25000 resamples were used to determine the rejection rates. It is further noted that where power curves appear incomplete (such as Figure 6 bottom left), the configuration of parameters led to occasional samples which consisted almost entirely of zeros, and hence could not be reliably fitted within the framework of a simulation study.

Figure 4: Attainment rate under hybrid estimator. Left: alternative $H_1^{(a)}$; right: $H_1^{(b)}$

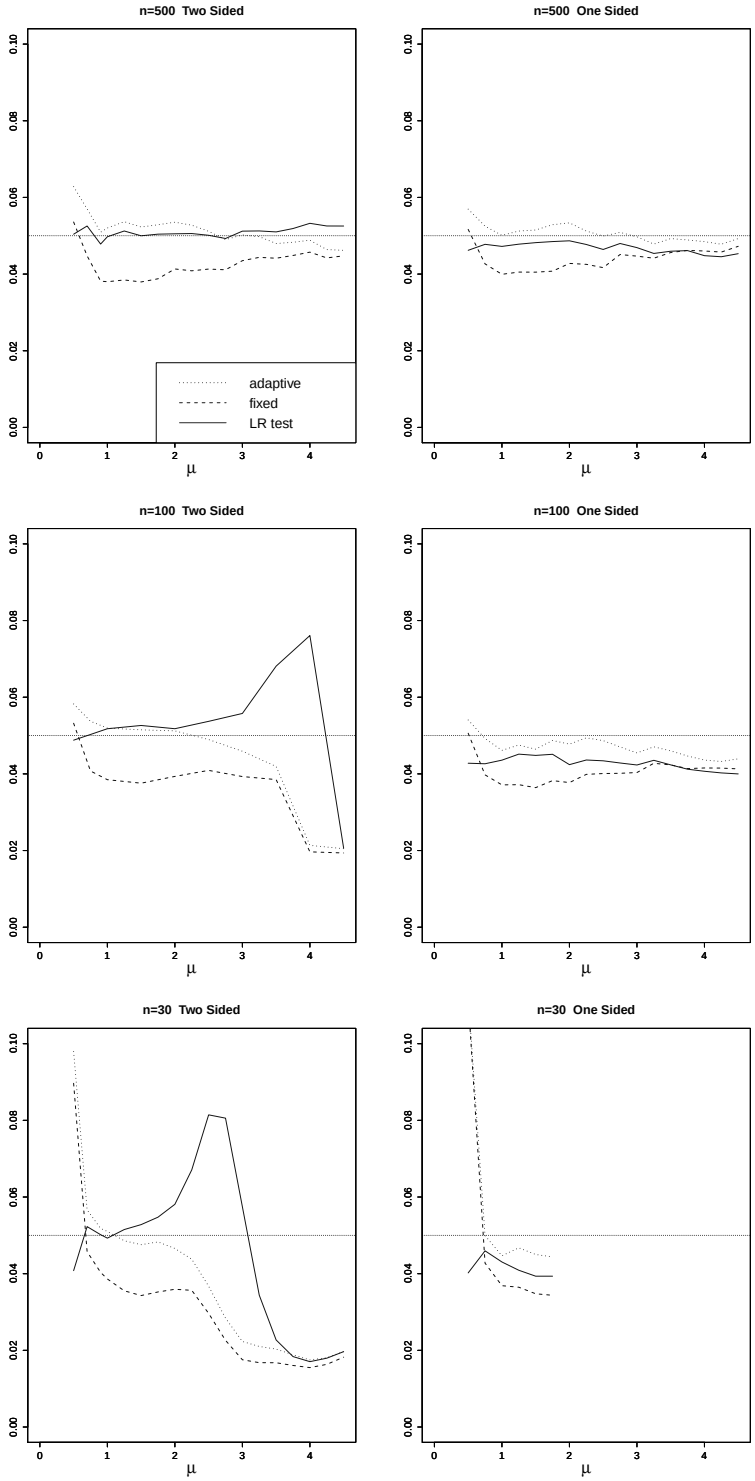


Figure 5: Power under hybrid estimators (test of zero–modification; $H_1^{(a)}$)

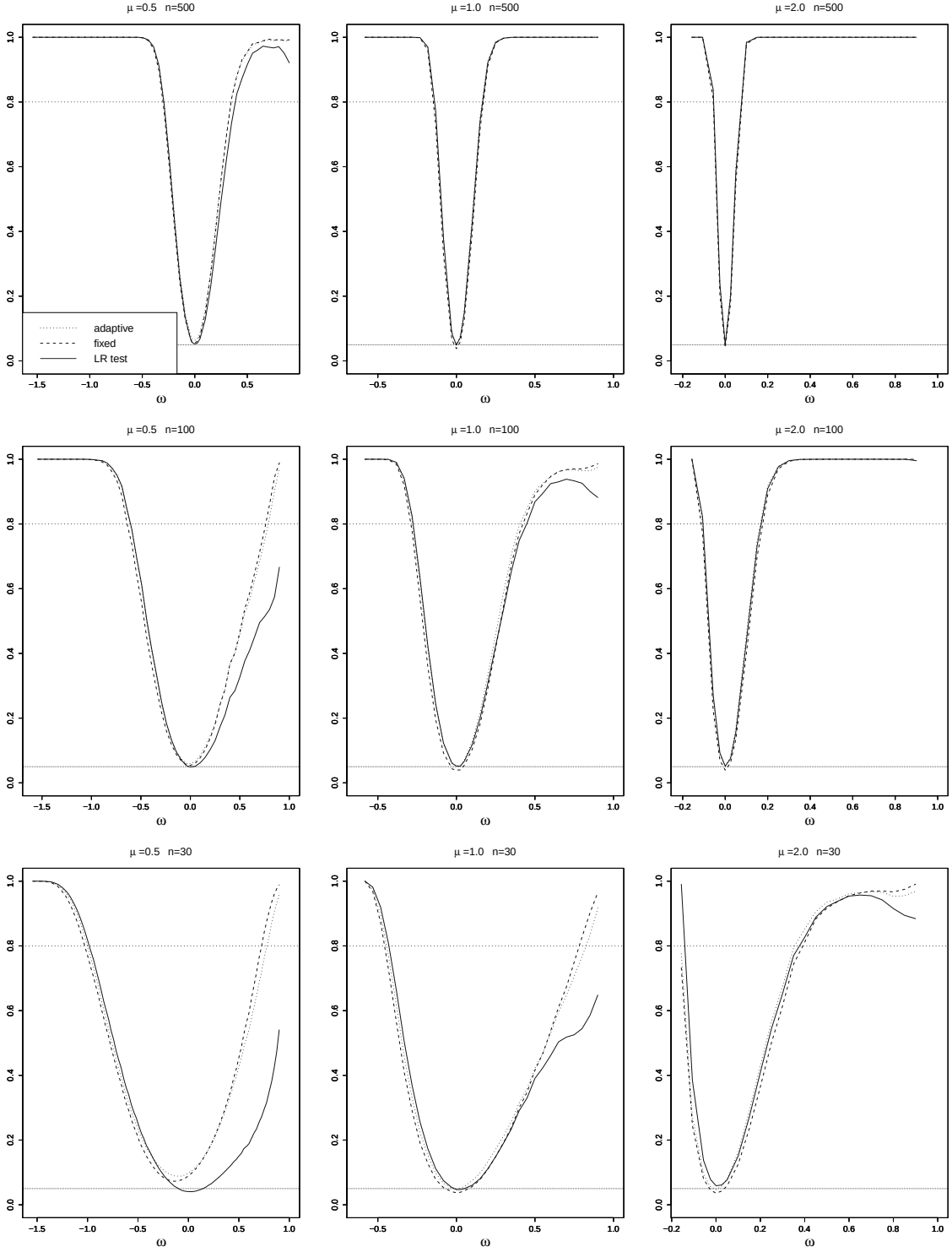
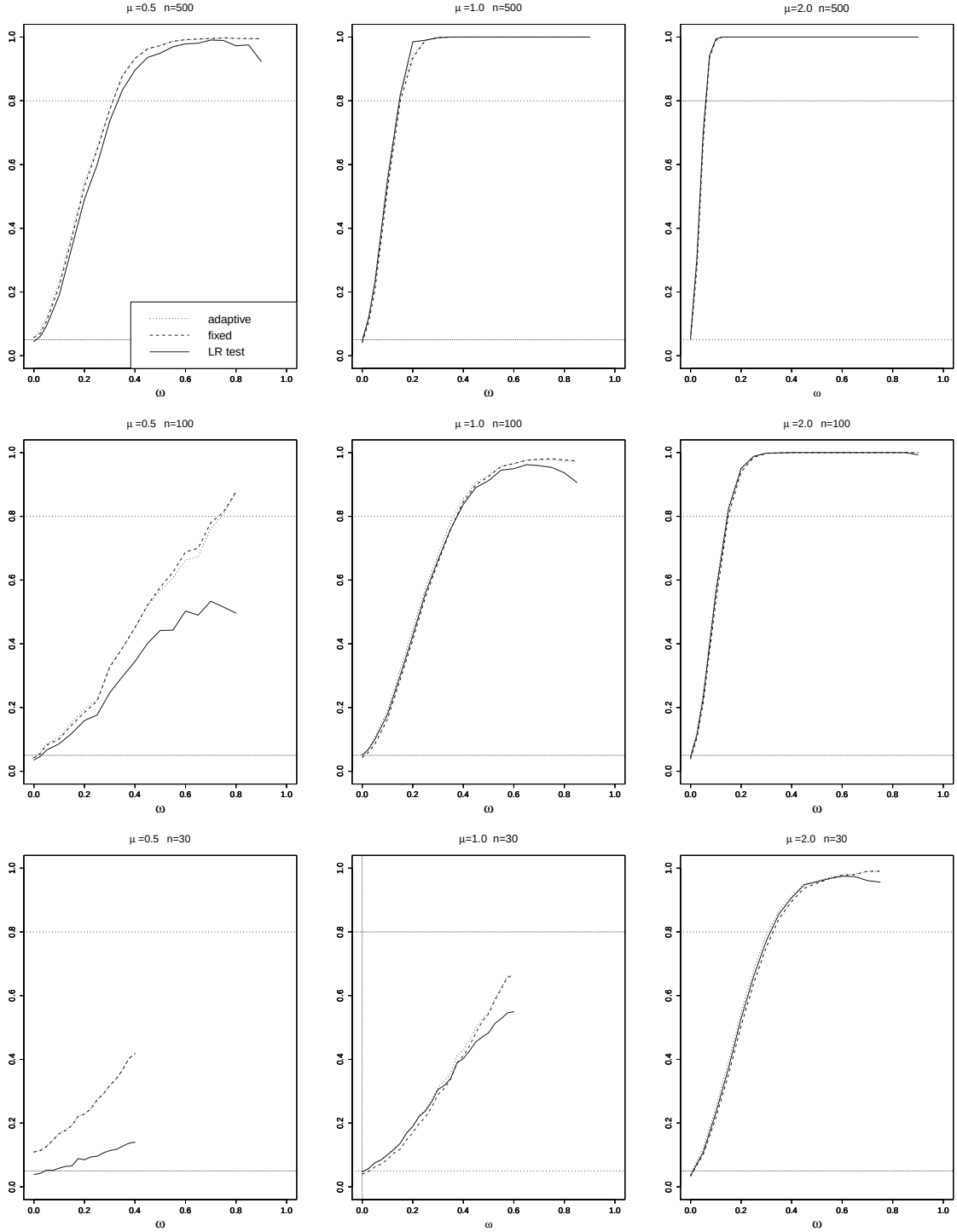


Figure 6: Power under hybrid estimators (test of zero–inflation; $H_1^{(b)}$)



4.5 Mean function estimation in the presence of covariates

In this case, the hybrid version of the estimated mean for case i may be obtained by computing fitted values under a Poisson model (say, $\hat{\mu}_{i,\mathrm{W}}$) and a zero–truncated Poisson model ($\hat{\mu}_{i,\mathrm{T}}$), respectively, and then applying the hybrid technique (4.7) on the respective pairs of fitted values. The fitted values from the ZTP model can be obtained using statistical software such as the R-package **VGAM** (Yee, 2010). [Of course, this methodology could also be applied in the absence of covariates, in which case the fitted values will all be equal.] Denoting by $\hat{\mu}_{i,\mathrm{H}}$ the resulting hybrid mean estimates, this implies that for scheme (i) one simply has

$$\hat{\mu}_{i,\mathrm{H}} = \frac{2}{3}\hat{\mu}_{i,\mathrm{W}} + \frac{1}{3}\hat{\mu}_{i,\mathrm{T}}, \qquad i = 1, \ldots, n, \tag{4.9}$$

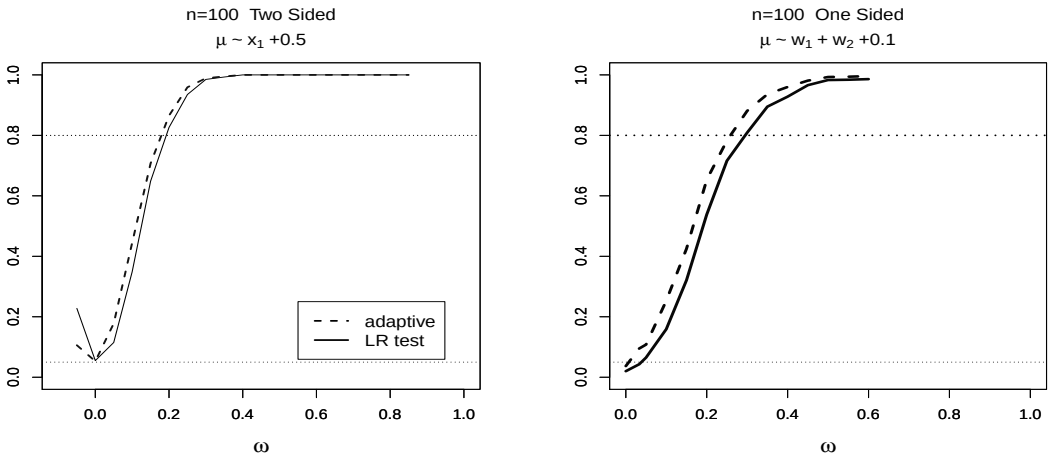
whereas for scheme (ii), an adaptive choice of h is obtained via $h_i = f(\hat{\mu}_{i,\mathrm{W}})$, yielding the case–wise hybrid rule

$$\hat{\mu}_{i,\mathrm{H}} = h_i\hat{\mu}_{i,\mathrm{W}} + (1 - h_i)\hat{\mu}_{i,\mathrm{T}}. \tag{4.10}$$

Figure 7 illustrates the power and attainment of the proposed test in comparison to the LR test in the presence of covariates. The left hand diagram displays the powers obtained when $n = 50$ observations are simulated from a zero–modified Poisson model, with zero modification parameter ω (on the horizontal axis) and Poisson parameter of the form $x_1 + 0.5$, where x_1 is a random draw of 50 observations from a uniform distribution on the interval $(0.5, 1.0)$. The right hand diagram illustrates the powers obtained when $n = 100$ observations are simulated from a zero–modified Poisson model with Poisson parameter of the form $w_1 + w_2 + 0.1$, where w_1 and w_2 are both random draws of 100 observations from a uniform distribution on $(0.2, 1)$. The adaptive hybrid parameter has been used, but the results remain similar for the fixed

estimator. Overall these plots give evidence that the proposed test compares strongly to the LR test also in the presence of covariates.

Figure 7: Power under hybrid estimator (covariate model, left: alternative $H_1^{(a)}$; right: $H_1^{(b)}$)



5 Examples

In this section we present a collection of examples, with and without covariates. R Code to reproduce these examples will be provided in the Statistical Modelling Archive under `www.statmod.org/smij/`.

We initially present an example of the proposed test applied to covariate–free data, in which case the Poisson–binomial distribution reduces to a binomial distribution, and proceed with two covariate–bearing examples in the subsections which follow. For all the examples of this section the adaptive (scheme (ii)) hybrid estimator of the Poisson mean was used in the execution of the proposed test.

For the one–sided tests of zero inflation the Poisson parameter is modelled by a

log link and the zero–inflation parameter by a logit link; for the one–sided tests of zero–deflation and the two–sided test of zero–modification both the Poisson and zero-inflation parameter are modelled by identity links. The estimates of the Poisson parameters necessary for the proposed test are derived from the Poisson and truncated Poisson models with log link for the one–sided tests, and with identity link for the two–sided tests.

5.1 The "Prussian horse kicks" data

Table 3: The "Prussian Horse Kicks Data"

y	0	1	2	3	4	≥ 5
count	144	91	32	11	2	0

Table 3 is the famous "Horse Kicks" data of von Bortkiewicz (1898), which summarises the number of deaths by horse or mule kicks per Prussian army corps annually between 1875 and 1894. Table 4 illustrates the use of one–sided and two–sided versions of the proposed test. Concerning the latter, we fail to reject H_0: *data is Poisson* in favour of $H_1^{(a)}$: *data is zero–modified Poisson* as the test statistic, i.e. the observed number of zeros, lies within the 95% MQI, or equivalently as $p = 0.30$. Note that this is in agreement with the results of a likelihood ratio test of the same hypothesis.

Further, we fail to reject H_0: *data is Poisson* in favour of the zero–inflated alternative $H_1^{(b)}$ since the observed number of zeros is not greater than the $95th$ quantile of the relevant binomial distribution (i.e. the upper limit of the 90% MQI, $n_{0.95}$), or equivalently as $p = 0.137$; similarly we fail to reject H_0: *data is Poisson* in favour of the zero–deflated alternative $H_1^{(c)}$ as the test statistic, i.e. the observed number of zeros is not less than the $5th$ quantile of the relevant binomial distribution (the lower

Table 4: One–and two–sided tests of zero–modification

		Proposed Test			LR Test	
H_1	n_0	95% MQI	90% MQI	p–value	Statistic	p–value
$H_1^{(a)}$	144	$[120.07, 148.23]$		0.30	1.026	0.288
$H_1^{(b)}$	144		$[118.09, 150.87]$	0.137	1.026	0.144
$H_1^{(c)}$	144			0.863		0.856

limit of the 90% MQI, $n_{0.05}$), or equivalently as $p = 0.856$. Again both these results are in agreement with the results of a likelihood ratio test of the same hypotheses.

5.2 Chromosome aberration data

We consider four datasets consisting of chromosome aberration counts in human blood cells after *in vitro* exposure to ionising radiation. These datasets have previously been studied by Oliveira et al. (2016), where detailed descriptions of the datasets can be found.

Table 5 summarises the results obtained when the proposed test and a likelihood ratio test are used to test for zero–inflation relative to a Poisson regression model with log link and a quadratic linear predictor for covariate 'absorbed radiation dose' [Gy]. The third column provides the 90% MQI, the upper bound of which coincides with the critical value $n_{0.95}$ for the zero–inflation test ($H_1^{(b)}$). We see that, for all data sets except A3, the observed number of zeros exceeds $n_{0.95}$, hence clearly rejecting the Poisson model in favour of the zero–inflated Poisson model for A1, B1 and C1, but not for A3. These results are in full agreement with the corresponding one–sided LR test. Note that the lower limit of the 90% MQI is included in Table 5 for informational

purposes, but is not required for test.

Table 5: Analyses of chromosome aberration data. Data labels refer to notation in Oliveira et al. (2016).

	Proposed Test $\left(H_1^{(b)}\right)$			LR Test	
Data	n_0	90% MQI	p-value	Statistic	p-value
A1	14 430	[14213.9, 14318.4]	$< 10^{-9}$	16.37	5.22×10^{-5}
A3	2 747	[2726.5, 2814.3]	0.368	0.98	0.322
B1	7 280	[6716.6, 6818.4]	$< 10^{-9}$	85.31	$< 10^{-9}$
C1	6 786	[5041.1, 5152.8]	$< 10^{-9}$	1330.65	$< 10^{-9}$

5.3 Trajan Data

The data are the number of roots produced by $n = 270$ micropropagated shoots of the columnar apple cultivar "Trajan". During the rooting period, all shoots were maintained under identical conditions, but the shoots themselves were cultured on media containing different concentrations of the cytokinin BAP, in growth cabinets with an 8 or 16 hour photoperiod. Full details of the experiment are to be found in Marin et al. (1993). A striking feature of the data is that although almost all of the 140 shoots produced under the 8 hour photoperiod rooted, only about half of the 130 shoots produced under the 16 hour photoperiod did. Overall 64 shoots produced zero roots, of which only 2 were from the shorter photoperiod.

These data were analysed by Ridout and Demétrio (1992) and Ridout et al. (1998). The latter paper presents a table of the fits of various Poisson and negative binomial models, and their zero–inflated counterparts, and finds evidence of zero–inflation with respect to both models. The authors comment that there is little evidence of an effect

due to BAP concentration, but the effect of photoperiod is significant.

Table 6: Poisson analyses of Trajan data.

	Proposed Test $\left(H_1^{(b)}\right)$			LR Test	
Data	n_0	90% MQI	p-value	Statistic	p-value
all	64	$[0, 4.9]$	$< 10^{-16}$	314.2	$< 10^{-16}$
period = 16	62	$[0, 4.13]$	$< 10^{-16}$	316.8	$< 10^{-16}$
period = 8	2	$[0, 0.55]$	0.003	7.846	0.003

The results when the proposed method is used with a Poisson model (where the mean is modelled by photoperiod) as the model of the null hypothesis are summarised in Table 6, noting again a very good agreement between the proposed and the LR test.

6 Conclusion

We have developed a novel test for zero–inflation or zero–deflation in count data models with or without covariates, which tackles the problem more directly than existing asymptotic tests, by asserting whether or not the observed number of zeros is plausible under the hypothesized count distribution. The plausibility is assessed with reference to appropriate quantiles of a Poisson–binomial distribution. Essential to this procedure is the estimation of the parameters of the count data model. The question of how to estimate the mean parameter robustly has been given detailed attention in the case of the Poisson hypothesis, and a 'hybrid' rule which mixes the whole sample mean with a zero–truncated mean estimator has been developed which yields excellent attainment and power properties of the resulting zero–modification test. This hybrid estimator was developed specifically for the purpose of the proposed

test, but may be of more general use than the one presented here. The extension of the test to other base distributions is straightforward, however the investigation of the requirement for, and shape of, robust parameter estimation techniques such as the hybrid estimator for other base distributions than Poisson requires further attention.

Acknowledgements

We would like to thank two anonymous referees for their insightful comments and suggestions which have led to a major improvement in the presentation and clarity of this manuscript.

References

Chen, S. X., and Liu, J.S. (1997) Statistical applications of the Poisson-binomial and conditional Bernoulli distributions. *Statistica Sinica*, **7**, 875 – 892.

Cox, D.R., (1962) Further results on tests of separate families of hypotheses. *Journal of the Royal Statistical Society Series B*, **24**, 406 – 423.

Dempster, A.P and Schatzoff, M. (1965) Expected significance level as a sensitivity index for test statistics. *Journal of the American Statistical Association*, **60**, 420 – 436.

Daskalakis, C., Diakonikolas, I. and Servedio, R.A (2012) Learning Poisson binomial distributions. *Proceedings of the 44th Symposium on Theory of Computing*, 709 – 728.

Dietz, E. and and Böhning, D. On estimation of the Poisson parameter in zero–modified Poisson models. *Computational Statistics & Data Analysis*, **34**, 547 – 548.

Franck, W. (1986) *P*–values for discrete test statistics. *Biometrical Journal*, **28**, 403 – 406.

Heimers, A., Brede, H.J., Giesen, U. and Hoffmann, W. (2006). Chromosome aberration analysis and the influence of mitotic delay after simulated partial–body exposure with high doses of sparsely and densely ionising radiation. *Radiation and Environmental Biophysics*, **45**, 45–54.

Hong, Y. (2013) On computing the distribution function for the Poisson binomial distribution. *Computational Statistics and Data Analysis*, **59**, 41 – 51

Hong, Y. (2013) poibin: The Poisson Binomial Distribution. url = http://CRAN.R-project.org/package=poibin

Irwin, J.O. (1959) On the estimation of the mean of a Poisson distribution from a sample with the zero class missing. *Biometrics*, **15**, 324 – 326.

Lancaster, H.O. (1961) Significance tests in discrete distributions. *Journal of the American Statistical Association*, **60**, 233 – 234.

Loeys, T., Moerkerke, B., De Smet, O., and Buysse, A. (2012) Expert tutorial: The analysis of zero–inflated count data: Beyond zero–inflated Poisson regression. *British Journal of Mathematical and Statistical Psychology*, **65**, 163 - 180.

Ma, Y., Genton M. and Parzen, E. (2011) Asymptotic properties of sample quantiles of discrete distributions. *Annals of the Institute of Statistical Mathematics*, **63**, 227 – 243.

Marin, J. A., Jones, O.P., and W. Hadlow W.C. (1993) Micropropagation of columnar apple trees. *Journal of Horticultural Science*, **68**, 289 – 297.

Mendonca, L., (1995) Longitudinalstudie zu kariespräventiven Methoden, durchgeführt bei 7– bis 10–jährigen urbanen Kindern in Belo Horizonte (Brasilien). *Inaugural–Dissertation zur Erlangung der zahnmedizinischen Doktorwürde am Fachbereich Zahn–, Mund– und Kieferheilkunde der Freien Universität Berlin.*

Molenberghs, G. and Verbeke, G. (2007). Likelihood Ratio, Score, and Wald tests in a constrained parameter space, *The American Statistician*, **61**, 22 – 27.

Oliveira, M., Einbeck, J., Higueras, M., Ainsbury, E., Puig, P. and Rothkamm, K. (2016) Zero–inflated regression models for radiation–induced chromosome aberration data: A comparative study. *Biometrical Journal*, **58**, 259 - 279.

Poisson, S.D. (1837) Recherches sur la probabilité des jugements en matière criminelle et en matière civile. Paris:Bachelier.

Plackett R.L. (1953) The truncated Poisson distribution. *Biometrics*, **9**, 485 – 488.

Ridout, M.S. and Demétrio C.B. (1992) Generalized linear models for positive count data. *Revista de Matemática e Estatística*, **10**, 139 – 148.

Ridout, M.S., Demétrio C.B. and Hinde, J. (1998) Models for count data with many zeros. *Proceedings of the XIXth International Biometric Conference*, **19**, 179 – 192.

Stone, M. (1969) The role of significance testing. Some data with a message. *Biometrika*, **56**, 485 – 493.

Vuong, Q. (1989) Likelihood ratio tests for model selection and non–nested hypotheses. *Econometrica*, **57**, 307 – 333.

von Bortkiewicz, L. (1898) *Das Gesetz der kleinen Zahlen*. Leipzig: BG Teubner.

Wilson, P. (2015). The misuse of the Vuong Test for non–nested models to test for zero–inflation. *Economics Letters*, **127**, 51 – 53.

Wilson, P. and Einbeck, J. (2017). Sample quantiles corresponding to mid p–values for zero–modification tests. *Proceedings of the 32nd International Workshop on Statistical Modelling*, Groningen, 275 – 279.

Yee, T.W. The VGAM package for categorical data analysis. *Journal of Statistical Software* **32**, 1 – 34.